REMARKS

Claims 1-4

Claims 1-4 were rejected under 35 U.S.C. § 102(a) as being anticipated by Smith et al. (U.S. Patent 6,408,271 B, hereinafter Smith).

Smith discloses a system for generating possible pronunciations of a sequence of words. Under Smith, each word has many possible pronunciations. As a result, for a sequence of words, there are multiple possible combinations of these pronunciations. Smith selects the top N pronunciations for the sequence of words to store in a dictionary and use during speech recognition. During speech recognition, a decoder compares input feature vectors to pronunciations in the dictionary to determine if any of the pronunciations match the user's speech.

Independent claim 1 provides a method of adding an acoustic description of a word to a speech recognition lexicon. Initially, the text of a word is converted into an orthographically derived acoustic description of the word. The orthographically derived acoustic description is then scored based in part on a comparison between the orthographically derived acoustic description and a speech signal representing a user's pronunciation of the word. The speech signal is also used to identify a speech-based acoustic description of the word and a score for the speech-based acoustic description wherein the speech-based acoustic description is not associated with the text of the word. One of the orthographically derived acoustic description and the speech-based acoustic description is then selected as the acoustic description of the word based on the scores for the two acoustic descriptions.

Smith does not show or suggest the invention of claim 1 because it does not show or suggest steps of identifying a speech-based acoustic description of a word from a speech signal representing a user's pronunciation of a word where the speech-

based acoustic description is not associated with the text of the word.

In particular, steps 202 and 302 in combination with speech signal 804 of Smith do not show this step.

In steps 202 and 302, an acoustic description is formed though text-to-phoneme rules or by using a transcription dictionary. (See Smith, col. 6, lines 16-20). When text-to-phoneme rules are used it is clear that the acoustic description is associated with the text of the word. When the transcription dictionary is used, the text of the word is applied to the dictionary to locate the acoustic description that has been associated with the text. Thus, in either case, Smith identifies an acoustic description that is associated with the text of the word.

This is substantially different from the invention of claim 1 where the acoustic description identified from the speech signal is not associated with the text of the word.

Claim 1 is further patentable over Smith because it includes a limitation to selecting between a speech-based acoustic description that is formed from a speech signal and an orthographically derived acoustic description that is formed from text. Smith does not show such a selection process. In particular the section at column 12, lines 26-37 of Smith does not show such a selection process. Instead, that section discusses generating possible pronunciations for a sequence of words based on possible pronunciations for the individual words in the sequence. All of these pronunciations are generated from text as stated in column 6, lines 16-20. Thus, none of the pronunciations are identified from a speech signal without using the text.

Since Smith does not show the steps of identifying a speech-based acoustic description from a speech signal where the acoustic description is not associated with the text of the word

or selecting between such a speech-based acoustic description and another acoustic description, it does not anticipate claim 1 or claims 2-4, which depend therefrom.

Claim 5

Claim 5 was rejected under 35 U.S.C. § 103(a) as being obvious from Smith in view of Bahl et al. (U.S. Patent 5,875,426, hereinafter Bahl '426).

Bahl '426 provides a speech recognition system that is able to handle word pronunciations that are context dependent. During recognition, Bahl '426 first considers all possible stored pronunciations for all words in a vocabulary. The speech signal is applied to these pronunciations to identify a set of candidate words. All of these pronunciations are associated with the text of the words. These candidate words are applied to a language model that generates a score for each current candidate word based on a previously identified word. This results in a ranked list of candidate current words and the dictionary-based pronunciations of those words.

Bahl '426 then examines a field in each current word's dictionary entry and a field in the preceding word's dictionary entry to determine if an additional pronunciation of the word should be added as a candidate. Note that this additional pronunciation candidate is a rule-based candidate associated with the text of the word and is not dependent on how the speaker pronounced the word. The speech signal is then applied to these candidate words and pronunciations in order to select a most likely word.

Dependent claim 5 depends from claim 1 and includes a further limitation wherein identifying a score for a speech-based acoustic description further comprises using a language model.

Because claim 5 depends from claim 1, it includes the limitation to identifying a speech-based acoustic description of a word by decoding a speech signal representing the user's

pronunciation of the word wherein the speech-based acoustic description is not associated with the text of the word. Neither Smith nor Bahl '426 show such a limitation.

This can be seen from the fact that neither Smith nor Bahl '426 is capable of producing an acoustic description without an acoustic description that is associated with the text of a word. In Smith, if the acoustic descriptions from the text-to-phoneme rules or transcription dictionaries are removed, there are no pronunciations to combine and thus no pronunciations to score during speech recognition. Similarly, if the acoustic descriptions of the words stored in the lexeme in Bahl '426 are removed, there would be no phonetic baseforms to use during fast match 102 and detailed match 106. Since neither Smith nor Bahl '426 can produce an acoustic description of a word without an acoustic description associated with the text, it is clear that neither is able to produce a speech-based acoustic description from a speech signal where the acoustic description is not associated with the text of the word. As such, claim 5 is patentable over the combination of Smith and Bahl '426.

Claims 6-8

Claims 6-8 were rejected under 35 U.S.C. § 103(a) as being obvious from Smith in view of Bahl '426 and further in view of Bahl et al. (U.S. Patent 6,377,921, hereinafter Bahl '921).

Bahl '921 provides a system for identifying transcription errors in text used for training a speech recognition system. Bahl '921 trains a set of acoustic models for acoustic units such as words, syllables, and phones. After the training is complete, a speech signal is aligned with its corresponding transcript using the trained models and a score is determined for each acoustic unit in the transcript. Instances of acoustic units that receive a low score from these models are then flagged and examined by a human operator to determine if the transcription is in error.

Claims 6-8 depend indirectly from claim 1. As a result, they include the limitation to identifying a speech-based acoustic description from a speech signal where the speech-based acoustic description is not associated with the text of the word. The combination of Smith, Bahl '426 and Bahl '921 does not show or suggest this limitation.

As discussed above, Smith and Bahl '426 fail to show this limitation. Similarly, Bahl '921 fails to show or suggest a step of identifying a speech-based acoustic description from a speech signal where the acoustic description is not associated with the text of the word. Under Bahl '921, the speech signal is applied to a known transcription of the speech that is associated with the text of the words. As such, Bahl '921 cannot identify an acoustic description that is not associated with the text of a word. Therefore, the combination of Bahl '921 with Smith and Bahl '426 does not show or suggest the invention of claims 6-8.

In addition, in claim 6, generating a score for a speech-based acoustic description includes generating a language model score for a sequence of syllable-like units. None of Smith, Bahl '426 or Bahl '921 show or suggest generating a language model score for a sequence of syllable-like units.

In the Final Office Action, language model 18B was cited as providing a language model score for syllable-like units. However, Bahl '921 never states that the language model uses syllable-like units. As such, it does not show or suggest generating a language model score for a sequence of syllable-like units.

Since none of the cited references show or suggest generating a language model score for a sequence of syllable-like units and since none of the cited references identify a speech-based acoustic description that is not associated with the text of a word, claim 6 and claims 7 and 8, which depend therefrom, are patentable over Smith, Bahl '426 and Bahl '921.

Claims 9-11

Claims 9-11 were rejected under 35 U.S.C. § 103(a) as being obvious from Smith in view of Bahl '426 and further in view of Contolini et al. (U.S. Patent 6,233,553, hereinafter Contolini).

Contolini provides a method of selecting one pronunciation from a set of text-based pronunciations. Under Contolini, a plurality of text-based pronunciations are formed from the spelling of a word using a transcription generator. The top N pronunciations are provided to a speech recognition system, which applies a speech signal to the transcriptions representing each pronunciation. The transcription that scores highest is selected for storage. Contolini does not identifying a speech-based acoustic description from a speech signal where the speech-based acoustic description is not associated with the text of a word, nor does it show the production of an acoustic model score for a syllable-like unit by generating acoustic model scores for each of a sequence of phonemes that form the syllable-like unit.

Claims 9-11 depend from claim 1 and as such include the limitation to identifying a speech-based acoustic description from a speech signal where the speech-based acoustic description is not associated with the text of a word. None of Smith, Bahl '426 and Contolini show or suggest such a step.

In Contolini, a speech signal is applied against previously identified transcriptions to identify a score for each transcription. Since each of these transcriptions is associated with the text of the word, Contolini does not identify a speech-based acoustic description that is not associated with the text of a word.

Since none of the cited references show a step of identifying a speech-based acoustic description from a speech signal where the acoustic description is not associated with the text of the word, claims 9-11 are patentable over the cited art.

In addition, none of Smith, Bahl '426 or Contolini show or suggest generating an acoustic model score for a sequence of syllable-like units by generating acoustic model scores for each of a sequence of phonemes that form the sequence of syllable-like units as found in claim 9.

In the Office Action, it was asserted that claim 4, column 7, line 6 and column 6, line 56 of Contolini show this limitation. Applicants respectfully dispute this assertion.

Claim 4 simply states that the sound units of claim 1 are acoustic units. Neither claim 1 nor claim 4 make any mention of syllable-like units or of determining an acoustic score for a syllable-like unit by determining acoustic scores for a sequence of phonemes that form the syllable-like units. Column 6, line 56 describes classes of phonemes including consonant and syllabic. This section does not suggest generating an acoustic score for a syllable-like unit by determining acoustic scores for a sequence of phonemes. Instead, it simply shows that a single phoneme may act as a syllable at times. When this occurs, forming an acoustic score for the syllable does not require determining the acoustic score for a sequence of phonemes. Instead, the acoustic score for a single phoneme is determined.

Column 7, line 6 discusses filtering unlikely sequences of phonemes. It does not show or suggest determining an acoustic score for a syllable-like unit by generating acoustic scores for each of a sequence of phonemes that form the syllable-like unit.

Since none of Smith, Bahl '426, or Contolini, show or suggest determining an acoustic score for a syllable-like unit by determining acoustic scores for a sequence of phonemes that forms the syllable-like unit, the combination of these references does not show or suggest claim 9.

Claims 12-17

Claims 12-17 were rejected under 35 U.S.C. § 102(a) as being anticipated by Gupta et al. (U.S. Patent 6,243,680 B1, hereinafter Gupta).

Gupta provides a system for selecting a pronunciation of a word for entry into a dictionary. Under Gupta, the text of a new word is first converted into a string of phonemes using a set of text-to-phoneme rules 412. These phonemes are placed in a graph structure with each branch in the structure being represented by a different phoneme. For each phoneme branch, a set of parallel branches are constructed, one for each phoneme that is similar to the initial phoneme in the graph. Additional parallel branches are then added for each allophone of each phoneme in the graph where an allophone is a particular pronunciation of a phoneme. Gupta then applies a set of speech utterances to the graph to score each path through the graph. The path with the highest score is selected as the pronunciation of the word.

Independent claim 12 provides a computer-readable medium having instructions for selecting a phonetic description of a word to add to a lexicon. These steps include receiving the text of the word and a speech signal representing a person's pronunciation of the word. The text of the word is converted into a text-based phonetic description while the speech signal is used to generate a speech-based phonetic description of the word without using the text of the word. Either the text-based phonetic description or the speech-based phonetic description is then selected for entry in the lexicon based on the correspondence between each phonetic description and the speech signal.

Gupta does not show or suggest the invention of claim 12 because it does not include a step of selecting between a

text-based phonetic description and a speech-based phonetic description.

In the background section of Gupta, different types of systems for identifying pronunciations of new words are described. In one system, an expert listens to the word and identifies the acoustic description. In a separate system, a continuous allophone recognizer is used that decodes speech utterances to identify an acoustic description that is not associated with a word. In another system, a set of text-based rules are used to form an acoustic description.

However, Gupta does not show or suggest determining an acoustic description from the text and an acoustic description from the speech signal and then selecting between the two acoustic descriptions. Instead, text-based acoustic descriptions are used in separate systems from speech-based acoustic descriptions.

Note that in the Gupta system itself, only text-based acoustic descriptions are used. Specifically, "[t]he feature vectors for each utterance are used to score the allophonic graph generated on the basis of the orthographic representation of the new word." (Gupta, col. 13, lines 61-63). Thus, graph scoring unit 404 does not generate a speech-based phonetic description that does not use the text of a word, but simply scores the text-based phonetic descriptions proposed by graph generator 400.

The fact that Gupta does not produce a speech-based phonetic description can be seen clearly by removing all of the phonetic descriptions that use text. If this is done, allophone graph generator 400 produces an empty graph because the graph is only populated using letter-to-phoneme rules. (see Col. 5, lines 24-39) This empty graph is provided to graph scorer 404, which is then unable to function since it does not have any phonetic sequences to apply the speech signal against. If Gupta produced a speech-based phonetic description, this would not be true since

the speech-based phonetic description would still be present even if the text-based phonetic descriptions were removed.

Since Gupta does not produce a speech-based phonetic description, it cannot select between a text-based phonetic description and a speech-based phonetic description. As such, it does not anticipate claim 12 or claims 13-17, which depend therefrom.

Claim 18

Claim 18 was rejected under 35 U.S.C. § 103(a) as being obvious from Gupta in view of Contolini.

Claim 18 depends from claim 12 and thus includes the limitation to generating a speech-based phonetic description of a word from a representation of a speech signal without using the text of the word. Neither Gupta nor Contolini show this limitation.

In particular, Contolini does not show or suggest producing a speech-based phonetic description from a speech signal without using the text of the word. Instead, Contolini simply applies a speech signal to previously defined phonetic descriptions in order to score each phonetic description. The phonetic descriptions are generated "based on spelled letter input, using a set of decision trees." (Contolini at Col. 3, lines 7-8).

In the Final Office Action, column 4, line 40 of Contolini was cited as showing that Contolini processes speech input. In the cited section, a user pronounces a single syllable of a word. Contolini then links that pronunciation to the original text spelling. Thus, in the cited section, Contolini uses the text of the word to deduce the location for the sound produced by the user. Without the text, Contolini would not be able to produce a phonetic description of the word from the speech signal.

This is substantially different from claims 12 and 18 where the speech-based phonetic description is generated from the representation of the speech signal without using the text of the word.

Since neither Gupta nor Contolini produce a speech-based phonetic description without using the text of a word, the combination of these two references does not show or suggest the invention of claim 18.

Claims 19-21

Claims 19-21 were rejected under 35 U.S.C. § 103(a) as being obvious from Schulze (U.S. Patent No. 6,167,369) in view of Gupta.

Schulze describes a system for determining the language of a document. To do this, Schulze generates a set of trigram models for each language, where each trigram model provides the probability of a character trigram in the language. An input text is then divided into trigrams. The trigrams for the input text are scored using the models for each language to generate a total score for each language. Schulze does not show or suggest syllable-like units or forming n-grams of syllable-like units.

Independent claim 19 provides a speech recognition system with a language model that is trained through a series of steps that include breaking each word in a dictionary into syllable-like units and for each word, grouping the syllable-like units into n-grams. The total number of n-gram occurrences in the dictionary is counted and for each n-gram, the total number of occurrences of the particular n-gram is divided by the total number of n-gram occurrences in the dictionary to form a language model probability for the n-gram.

The combination of Schulze and Gupta does not show or suggest the invention of claim 19. In particular, neither reference shows or suggests grouping syllable-like units found in dictionary words into n-grams.

In the Office Action, it was asserted that Schulze shows grouping syllable-like units from dictionary words into n-grams at column 1, line 29. Applicants respectfully dispute this assertion.

The cited section of Schulze discusses dividing an input sentence into individual character trigrams. It does not mention syllable-like units or forming n-grams from syllable-like units. Furthermore, it would not be obvious to use syllable-like units with Schulze. One goal of the Schulze system is to be able to identify the language of short text segments. If larger units were used instead of individual characters, there would be fewer n-gram probabilities calculated for short text segments thereby making it more difficult to identify the language of the text.

In the Final Office Action, it was asserted that the sub-word units of Gupta correspond to a syllable-like unit. Thus, the rejection appears to be based on substituting the sub-word units of Gupta in the technique described by Schulze.

However, those skilled in the art would not make such a substitution. Under Schulze, the language of the text is unknown. Because of this, it would be very difficult and in some cases may be impossible to divide the words into syllable-like units. In fact, for some languages in Schulze, the text cannot even be divided into words. (See Schulze Col. 15, lines 50-53). Thus, those skilled in the art would not apply the sub-words of Gupta to Schulze as suggested by the Examiner. As such, claim 19 and claims 20 and 21, which depend therefrom are patentable over the combination of Gupta and Schulze.

Claims 20 and 21 are additionally patentable over Schulze and Gupta. In claim 20, the dictionary words are broken into syllable-like units by preferring syllable-like units that occur more frequently in the dictionary than other syllable-like units. Neither Schulze nor Gupta show or suggest this additional limitation.

In the Office Action, it was asserted that Schulze showed preferring syllable-like units that occur more often at column 12, lines 35-37. However, the cited section does not discuss syllable-like units or providing a preference for certain speech units when dividing a dictionary word into speech units. Instead, the cited section states that trigrams with low frequency counts are discarded from a trigram array.

Trigrams found in a corpus cannot be given a preference during the search for the trigrams. The reason for this is that there is no latitude in how trigrams are identified in a word. Under Schulze, the trigrams are identified simply by selecting three characters in a row in a word. Just because one three-character sequence is later removed from the array does not influence the identification of the trigrams in the words. All of the trigrams are identified regardless of which ones are later discarded from the array.

Since the rules for identifying trigrams do not allow a preference to be applied so that one trigram is preferred over another during trigram identification, the cited section of Schulze cannot show or suggest preferring syllable-like units that occur more often in a dictionary over other syllable-like units when dividing words into syllable-like units. As such, the combination of Gupta and Schulze does not show or suggest the invention of claims 20 and 21.

Conclusion

In light of the above remarks, claims 1-21 are patentable over the cited art. Reconsideration and allowance of the claims is respectfully requested.

The Director is authorized to charge any fee deficiency required by this paper or credit any overpayment to Deposit Account No. 23-1123.

Respectfully submitted,

WESTMAN, CHAMPLIN & KELLY, P.A.

By: _____

Theodore M. Magee, Reg. No. 39,758
Suite 1600 - International Centre
900 Second Avenue South
Minneapolis, Minnesota 55402-3319
Phone: (612) 334-3222 Fax: (612) 334-3312

tmm